Georgia Institute of Technology Technical Report : GIT-GVU-TR-01-16

Visual Coding and Tracking of Speech Related Facial Motion

Lionel Reveret[1,2]

[1]iMAGIS-GRAVIR
INRIA Rhône-Alpes
Montbonnot, FRANCE

Irfan Essa[2]

[2]CPL-GVU Center / College of Computing
Georgia Institute of Technology
Atlanta, USA

Abstract

This article present a visual characterization of facial motions inherent with speaking. We propose a set of four Facial Speech Parameters (FSP): jaw opening, lips rounding, lips closure, and lips raising, to represent the primary visual gestures of speech articulation into a multidimensional linear manifold. This manifold is initially generated as a statistical model, obtained by analyzing accurate 3D data of a reference human subject. The FSP are then associated to the linear modes of this statistical model, resulting in a 3D parametric facial mesh. We have tested the speaker-independent hypothesis of this manifold with a model-based video tracking task applied on different subjects. Firstly, the parametric model is adapted and aligned to a subject's face for a single shape. Then the face motion is tracked by optimally aligning the incoming video frames with the face model, textured with the first image, and deformed by varying the FSP, head rotations, and translations. We show results of the tracking for different subjects using our method. Finally, we demonstrate the facial activity encoding into the four FSP values to represent speaker-independent phonetic information.

1 Introduction

Several recent works in vision and machine learning showed that linear manifolds can be efficiently applied to model variations of facial morphology. Blanz and Vetter showed a morphable model of morphological variations, learned from a large database of 3D facial scan [1]. The Active Appearance Models of Cootes et al. showed similarly an efficient coding of facial shape and texture by a linear model [5]. Pighin shows that a linear combination of canonical facial expressions (joy, surprise, etc.) can be used to track intermediate expressions from video [7].

The motivation of our work is to show that a similar approach can be proposed for speech movements, which can be modelled as a multi linear manifold of shape and appearance. This manifold is driven by four parameters, that we introduce as Facial Speech Parameters (FSP): jaw opening, lip rounding, lip closure, and lip raising. These parameters represent the primary visual gestures of speech articulation. We learn this manifold from a statistical method applied on a reference subject, carefully labelled with 3D markers.

In the context of face-to-face communication, speech is more than the transmission of an acoustical signal. The production of speech sounds is related to very specific and stable geometrical configurations of the lips and the jaw. Human beings are constantly exposed to both the acoustical stimuli and their visual correlates on the face. We perceive, and are very sensitive to, the spatio-temporal coherence between the sounds of speech and the facial gestures that are served to partially "shape" those sounds [4, 32]. Even animations of talking faces are subjected to this ontological fact in order to convey a believable perception [25]. Like sounds, these gestures appears similar from one subject to an other, despite the change of morphology, and have to be shared as a common basis for audiovisual communication [4]. This observation motivated to test the speaker-independent properties of the FSP manifold learned on the reference subject. This hypothesis is similar to the modelling of DeCarlo and Metaxas [6] in the sense that morphology and gesture are separately parameterized. In different, but related works, this hypothesis of stability of visual representations of speech has been used in different speaker-dependent cases to predict facial motion from acoustic [3, 2]. We propose here a parametric characterization of this visual stability of speech in a speaker-independent perspective, i.e. invariant for different subjects.

The complexity of the non-rigid deformation of facial movements coupled with the lack of robust features motivates the use of parameterized motion models to regularize the automatic analysis of face images. Such models have been used to track head movements, recognize expressions, and measure details up to the level of quantifying eyebrow raises and lip curls. Only a few simple attempts have addressed the coding and the automatic analysis of speech motion. This can be partly attributed to the lack of an existing experimental specification similar to the well-established Facial Action Coding System (FACS) proposed by Ekman and Friesen [16] for encoding motion of facial expressions. Previous works in automatic lip reading have attempted the recognition of a closed vocabulary (letters, digits, isolated words) or features tracking [14, 24, 28], but rarely a robust and high-level motion recovery has been addressed as it has been for expressions [6, 17, 13]. Some recent work on coding facial motions, called Facial Animation Parameters (FAPs), which are now included in the MPEG4 specifications, do model lip and mouth shapes [35, 31, 15]. However, FAP focus more on the coding of low-level features, rather than more constrained and meaningful gesture, such as a smiling gesture. Some of the above-mentioned contributions and a survey report [27] suggest the importance of building an encoding system that is more suitable for modelling visual speech. This is specifically the goal of our research effort.

In this paper, we show that an accurate model of speech motion learned from data of an expert subject can be re-used to track other subjects' face motion after a morphological (geometric, structural) adaptation. This model implements a high-level encoding of speech motion, which aids in the automatic visual recognition of non limited vocabulary (*i.e.,* not restricted to the learning set), as well as photo-realistic and non photo-realistic facial animation. We demonstrate the ability of this model to encode visual speech action parameters from video tracking of lips and face motion of talking subjects. We demonstrate its capabilities to be used for tracking long sequences of lip and face movements in a model-based approach. We improve on the tracking by adding the a texture-based approach, which provide increases robustness.

2 Related Work

There is considerable work in the area of face processing from video. Most of it concentrated on model-based tracking of face movement. Here we undertake a brief exposition of this motivating work that aids in the development of our specific model for speech gestures. We specifically concentrate on earlier works on model-based tracking and on methods for registration of model from texture alignment.

Model-based tracking: DeCarlo and Metaxas [6] have successfully demonstrated the use of a parameterized face model to track movement of the head, smiling and mouth opening for different subjects. This approach uses a hand-designed model of face motion, with one single control parameter for the opening of the mouth. The DeCarlo and Metaxas method of tracking adds a stronger model to extend the Black and Yacoob [13] approach, where simple affine motion models were used to measure deformations. Black and Yacoob relied on FACS model to model facial expression. In the case of speech production, lips and face deform in a complex way, which cannot be represented with only one degree of freedom. The authors mention a need for better parametric representation of speech motion in their paper.

Physical models of faces have been proposed for analysis of the facial motion as they allow for more degrees of freedom [17, 34]. However, the modelling is mainly focused on solving the tracking of canonical facial expression and does not, at present, model the motion of speech production.

Basu *et al.* [11] have addressed the motion of lip movement in speech production. In this work, a model of lip motion is learned from video for each subject from the tracking of ink markers on the lip surface. After the learning phase, this model allows for accurate tracking of outer lip contours feature, but does not implement a general coding of lip motion. About 10 degrees of freedom are necessary for each subject, which could result in instability in the optimization procedure.

Some recent techniques on analysis and synthesis of faces with speech have shown significant promise. For example, Video Rewrite [3] generates facial animations by reordering existing video frames. The choice of frames to play is determined by analyzing the audio track to extract phonemic information and its relationship to training video data. Voice Puppetry [2] is yet another impressive technique that generates facial motion using the raw audio signal. It achieves this by learning a facial control model by analyzing video and audio of real facial behavior, automatically incorporating vocal and facial dynamics such as co-articulation. Both the Video Rewrite and Voice Puppetry techniques are however bound by needs of extensive data, with a related training phase on acoustical signal. We base our approach on the hypothesis that the *morphological variability of facial motion between different speakers is easier to solve than acoustical normalization.*

Model registration from images: Traditionally, optical flow has been used to provide pixel level information to align motion of the model onto the image brightness flow. The model-based approach in this case consists of regularizing the brightness consistency equation of the optical flow, into a model-based formulation from the a priori model of the face movement [6, 17, 13, 12, 23].

Some approaches show that a texture-based formulation of object tracking can be proposed as an alternative to optical flow. The Active Blob technique [10] implements a texture-based tracking of any deformable object with closed boundaries. Using statistical modeling of shape and texture, Active Appearance Models [5, 7] have been used to model and register differences in facial morphology. The

texture presents a higher robustness than optical flow, as it is not subjected to error accumulation [30, 21, 20]. In addition, from the perspective of real-time implementation, recent developments in 3D graphic hardware for texture rendering, makes available high performance texture-mapping at a low cost. We implemented a similar approach for the registration of our specific model from each image of a video sequence.

3 Learning of a reference FSP manifold

3.1 3D data collection

Laser range-finding methods (e.g., Cyberscan) have been widely used for the registration of 3D face structure. Although this method delivers dense information (typically few thousands polygons meshes), it does not provide an easy point-to-point matching between a set of different scans, introducing significant noise variability for statistical analysis. Some systems of active markers provide accurate "flesh-point" 3D data (e.g., Optotrack) but usually only a limited number of markers is available and their size makes them impossible to be set on lips. In our case, we used video analysis of passive markers set on the subject's face and 3D stereoscopic reconstruction. 148 spherical plastic beads (ø2mm) were glued on the face. The subject was filmed under 5 different viewpoints so that any bead could be always seen on 2 cameras at least. The 5 cameras have been calibrated using a special 3D object with known dimensions. For each camera, an optimization procedure extracted the extrinsic (3D position with the regard to the calibration object) and intrinsic parameters (perspective projection on the image plane). The spatial resolution of every camera was approximately 0.5mm per pixel. The 3D stereoscopic reconstruction consists of: (1) finding the correspondent 2D locations of a marker on different views, and (2) optimizing the 3D coordinates for the marker so that its projection on the image plans of the camera best fit the expected 2D locations. A maximum projection error of 4 pixels has been observed, what we consider acceptable as it corresponds to the size of a bead.

As no beads were glued on the lip surface, the lip shape has been registered using the 3D geometric model described in [8]. This 3D lip model is controlled by the location of 30 control points. A cubic polynomial surface interpolates these 30 points as follows: initially, 3 groups of 10 points define three basic contours (outer contour, inner contour and an intermediate contour) which are defined as piece-wise continuous cubic curve. Then, these 3 contours are interpolated by an orthogonal set of 10 cubic curves to shape the surface; specific geometrical rules constrain the XYZ coordinates of the points to resolve ambiguities in their setting.

Finally, the face of the subject has been registered with 178 points in 3D for every image. A hand labelling of the markers on the different views allows the extraction of the 3D location of these points.

3.2 Building the model

The analysis method described in [9] is based on iterative applications of Principal Component Analysis (PCA) on subset of points, specific to the jaw and the lips. Then, linear regressions between the 3D face data and the resulting components are used as linear predictor to evaluate the correlation between facial motion and the specific articulators. These components (here the Facial Speech Parameters, FSP) serve as linear parameters to control the 3D facial motion. Head motion must be removed between

different so that, only the variation due to speech movement is analyzed. The learning set of facial shapes consists of 34 key shapes selected from a phonetically balanced corpus (i.e., balanced coverage of phonemes realization), hand labelled to provide the 3D locations of the face markers. The following procedure provides four FSP from the set of 3D face data.

Let $X_i = [x_{ij}, y_{ij}, z_{ij}]_{j=1..p}$ be one of the $n = 24$ shape vector of $p = 178$ vertices of the 3D face model and $X = [X_i^t]_{i=1..n}^t$ be the matrix collecting the set of $n = 34$ learning shapes X_i.

Let $X_J = [x_{ij}]_{i=1..n, j \in I_J}$, $X_L = [x_{ij}]_{i=1..n, j \in I_L}$ and $X_R = [x_{ij}]_{i=1..n, j \in I_R}$ be the subsets of points from X, respectively "jaw", "lips" and the remaining points, forming a partition of X.

- the 3D data are centered on the mean shape over the training set :

$$\mu = \frac{1}{n} \sum_{i=1}^{34} X_i^{(0)} \tag{1}$$

$$X_i := X_i - \mu, \forall i = 1..n \tag{2}$$

- a PCA is applied on the "jaw" subset of points. The parameter value of the first component C_J is taken as the first FSP values over the training set;

$$C_J = argmin_C \|X_J - CE_J^t\|^2 = E_J X_J \tag{3}$$

  where E_J is the first eigenvector of the covariance matrix of X_J.

- a linear regression between these values and the 3D face data gives the first linear mode Φ_J associated with this FSP.

$$\Phi_J = argmin_\Phi \|X - C_J \Phi\|^2 = (C_J^t C_J)^{-1} C_J^t X \tag{4}$$

  Note that E_J and Φ_J are different due to the fact that E_J are the eigenvectors from a PCA on a subset of points and Φ_J are computed from a linear regression on the whole set points. Φ_J can be considered as the linear coupling of the whole facial motion with the principal motion of the jaw given by E_J.

- the contribution of the first parameter is removed from the original 3D data set;

$$X := X - C_J \Phi_J \tag{5}$$

- a PCA is applied on the "lips" subset of points. The parameter value of the three first components C_L are taken as the three following FSP value over the training set;

$$C_L = argmin_C \|X_L - CE_L^t\|^2 = E_L X_L \tag{6}$$

  where E_L is the matrix of the first three eigenvectors (column vector) of the covariance matrix of X_L.

- a linear regression between these value and the 3D face data gives the three last linear mode Φ_L associated with this FSP.

$$\Phi_L = argmin_\Phi \|X - C_L \Phi\|^2 = (C_L^t C_L)^{-1} C_L^t X \tag{7}$$

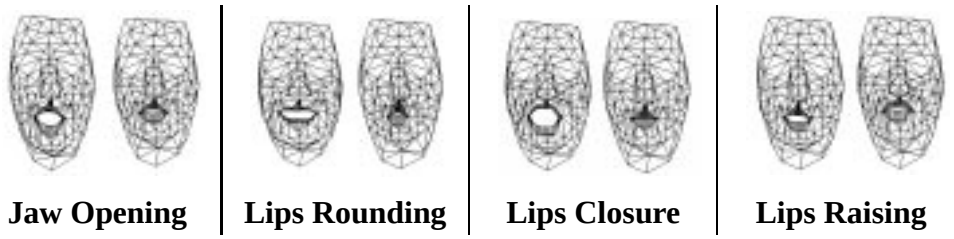

Figure 1: Our resulting 3D model and the extreme variations along the 4 FSPs (± 3 standard deviations from the mean shape). Opening of the jaw (FSP1), rounding of the lips (FSP2), closure of the lips (FSP3), raising of the lips (FSP4).

Finally, putting the four modes into a single matrix Φ, this spatial model of facial action allows us to generate a 3D model with actions represented as a linear combination of the 4 modes, $\Phi_1, \Phi_2, \Phi_3, \Phi_4$, controlled by parameters a that we introduced as Facial Speech Parameters (FSP).

Consider $X(a) = [x_j(a), y_j(a), z_j(a)]_{j=1..p} \in \Re^{3p}$ that describes a 3D geometric model, and μ as its mean shape, then we have a deformation model, which can be controlled by varying the FSPs (a) :

$$X(a) = \mu + \sum_{i=1}^{4} a_i \Phi_i = \mu + \Phi a, \tag{8}$$

These modes have been interpreted as phonetically pertinent gestures, consistent with [9]: (1) the opening the jaw; (2) the lips rounding, used to separate rounded vowel like [u] and spread vowel like [i]; (3) the closure of the lips for bilabial stop consonant like [p] [b] [m]; (4) lips raising for labio-dental fricatives consonants like [f] [v].

The study in [9] exhibits two more modes : (5) advance of the jaw; and (6) a remaining motion due to the raising of the pharynx. Although the last two parameters listed above are present in this new subject as well, they have very limited variations, especially from the frontal views. In practice, we also observed that these last two parameters resulted in some instability for the automatic estimation from video. Consequently, we have chosen to ignore them for most of the analysis in this paper and focused on the automatic extraction of the first four FSP parameters.

The Figure 1 shows our resulting 3D model and the extreme variations along the 4 FSPs (± 3 standard deviations from the mean shape).

3.3 Aligning the model to different subjects

The procedure described above provides a detailed model of facial movements at the cost of a time-consuming hand labelling of markers on several shapes. To reduce this step, we introduce a method to align the morphology of the initial model on a new subject, while keeping the same description of motion learned from the reference subject. This hypothesis is based on the observation that, despite difference in morphology, any human vocal track is subjected to the same spatial constraints and therefore will deform in a similar way, including for the face, motion of the jaw and lips. This hypothesis is similar to the modelling of DeCarlo and Metaxas [6] in the sense that morphology and gesture are separately parameterized. In our case, we benefit from a detailed model of facial deformation for speech,

learned from real data of a human subject. Our normalization can be formulated as an update of the mean shape in equation 8, the remaining FSP modes being kept identical for the new subject. This operation can be seen as a pure translation of the FSP manifold, initially learned from the reference subject. Having,

$$X_{ref}(a) = \mu_{ref} + \sum_{i=1}^{4} a_i \Phi_i, \tag{9}$$

from 8 for the reference subject, we model the new subject as,

$$X_{new}(a) = \mu_{new} + \sum_{i=1}^{4} a_i \Phi_i \tag{10}$$

whose 3D mesh is controlled by the same FSP parameters. This normalization implies to find the numerical values of the new mean shape μ_{new}.

To process this update, the 3D model of the reference subject in a rest position is wrapped onto the face of the new subject based on features alignment. These features are represented into a simplified model of the face with 20 nodes, with each node corresponding to a specific node in the original 3D model. This local alignment of the mesh vertices is done similarly to [7]. Once reference features are put in correspondence, a Radial Basis Functions (RBF) relaxation of the displacements interpolates the displacement of the remaining vertices not covered in the simplified mesh.

The displacement of the 3D mesh in rest position between the reference subject $X_{ref,rest}$ and the new subject $X_{new,rest}$ is noted as a function $D(X_{ref,rest})$ of the reference mesh. Each point $X_{new,rest,j} = [x_j y_j z_j]^t_{j=1..p}$ is given as :

$$X_{new,rest,j} = X_{ref,rest,j} + D(X_{ref,rest,j}), \forall j = 1..p \tag{11}$$

For a subset of 20 points , corresponding to the simplified model, the features $F_j = [u_j v_j]_{j \in I_P}$ are hand-labelled on the image of the new subject in a rest position. The calibration of the camera is assumed to be known. A rough approximation of the calibration is enough : even if the proportions of the model will not match the real proportions of the new subject, the 2D projection of the 3D model will fit the subject's image and thus later, allows the registration of the FSP parameters from images.

When only one view is available, the z component (orthogonal to the camera plane) is set to be equal between the vertices of the reference subject and the new subject. The initial orientation of the head is set to face the camera in front view - no rotation. This leads to an estimation of the 3D position of the 20 points $X_j = [x_j y_j z_j]_{j \in I_P}$ of the reduced face model on the new subject from the 2D features points F_j. Thus, the displacements $D_(j)$ between the reference subject and the new subject are known for this rest position :

$$D_j = X_{new,rest,j} - X_{ref,rest,j} = D(X_{ref,rest,j}) \forall j \in I_P \tag{12}$$

The displacement $D(X \in \Re^3)$ is expressed as a RBF interpolation function from $\Re^3$ to $\Re^3$:

$$D(X) = \sum_{j \in I_P} h(\|X - X_{ref,rest,j}\|) C_j \tag{13}$$

where C_j are $\Re^3$ vectors to be determined and $h(r)$ a RBF function. For this function, we choose :

$$h(r) = \exp^{-r^2/\alpha} \tag{14}$$

Figure 2: The original model for the reference model; the hand label features set on the new subject; the result of the original model aligned on the specified features.

The value of α has been determined so that the value of $h(r)$ is equal to 0.5 at a distance of 5 mm from its center (e.g., zero).

With $C = [C_j^t]_{j \in I_P}$ and $H(X) = [h(\|X - X_{ref,rest,j}\|)]_{j \in I_P}$, the equation (13) can be formulated in a matrix form as :

$$D(X) = H(X)C \tag{15}$$

Building the matrices $D = [D_j]_{j \in I_P}^t$, and $H = [H(X_{ref,rest,j})^t]_{j \in I_P}^t$, the interpolation coefficients in C are the solution of :

$$D = HC \tag{16}$$

Finally, for every n points of the mesh, the rest shape of the new subject is obtain from the reference subject as :

$$X_{new,rest,j} = X_{ref,rest,j} + H(X_{ref,rest,j}H^{-1}D, \forall j = 1..p \tag{17}$$

Figure 2 shows the results of the alignment of the reference model to a new subject.

For both subjects, reference and new, the rest position is coded by the same FSP numerical configuration a_{rest}, which sets the initial model into a position where jaw and lips are closed (FSP_1, FSP_3), with a neutral spreading of the lips (FSP_2) and no raising of the upper lip (FSP_4).

Consequently, given $X_{ref}(a_{rest})$ being the rest position $X_{ref,rest}$ of the reference subject and $X_{new,rest}$ being the result of the morphologic adaptation process described above for the new subject in a similar rest position, we obtain the mean μ_{new} for the new subject simply by subtraction. Per Equation (10), we have

$$X_{new}(a_{rest}) = \mu_{new} + \Phi a_{rest}. \tag{18}$$

Then for $X_{new,rest} = X_{new}(a_{rest})$, we get

$$\mu_{new} = X_{new,rest} - \Phi a_{rest}. \tag{19}$$

To validate the articulatory hypothesis (*i.e.*, the usage of the same FSP modes for different subjects), we hand labelled the simplified face model for four different speakers, doing 6 facial configuration, while uttering [a], [i], [u], [p] in [apa] and [f] in [afa]. The FSP configuration is recovered from the features location by an optimizing procedure that minimize the euclidian distance between the 20 hand-labelled features and the 2D front view projection of the corresponding model points driven by (8).

We obtain the following results, showing a pertinent repartition of the shapes according to the FSP interpretation even after the morphological adaptation (Figure 3). The jaw parameter (FSP1) isolate the [a] shape (wide opening), the protrusion parameter (FSP2) separates [u] shapes, the lips closure parameter (FSP3) separate the vowels from consonants that require a joining of the lips and finally the lip raising parameter (FSP4) separate the [p] consonants from the [f] consonants.

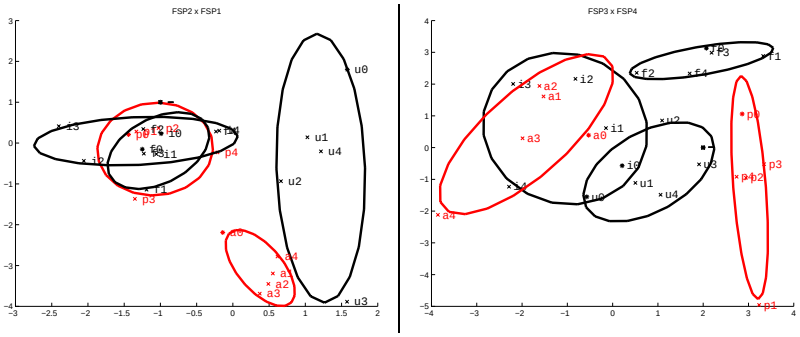

Figure 3: Plots showing variations in FSP amongst our subjects. The value of the FSP parameters of the reference speaker (index 0) are represented with stars, while the other 4 new speakers are represented with crosses (indexes 1 to 4). They are identical for the rest position (label [-]) by construction (see section 3) and differ for all the other shapes.

4 FSP registration from texture

4.1 Objective function

Compared to the approaches on optical flow in model-based tracking [6], registration from texture presents an interesting alternative as it is not heavily penalized by risk of drift. For any new face, taking the initial image for the morphological alignment allows us to set a texture correspondence for the model. Now, the texture follows subsequent deformation of the model and provide a synthetic image of the face. The tracking consists of finding the four numerical value of the FSP, plus translation and rotation that minimizes the difference between the projection of the textured model and the incoming image to analyze. We present this mathematically as follows:

For a set of position, rotations, and FSP parameters, we introduce

$$p = [r_x, r_y, r_z, t_x, t_y, t_z, a_1, a_2, a_3, a_4] = [p_j]_{j=1..10}, \tag{20}$$

I the RGB image to analyze and $\hat{I}(p)$ the image synthesized by the textured model. $\|\dots\|$ is the Euclidean norm on the RGB components of the image. The objective function is defined as follow :

$$e_i(p) = \hat{I}(p)(u_i, v_i) - I(u_i, v_i) \tag{21}$$

$$E(p) = \frac{1}{n} \sum_{i=1}^{n} \|e_i(p)\|^2 \tag{22}$$

where (u_i, v_i) defines the screen position of the pixel i in the image and in the original image, covered by the textured model. The n pixels considered are only those covered by the model projection. This screen area is extracted using the rendering buffer of OpenGL while displaying the textured model. The objective function is minimized with a Levenberg-Marquaardt optimization. Head position is recovered as well as the FSP parameters. However, in our test sequence, the subjects were asked to constantly look at the camera to limit the head motion. Robust tracking of head motion has been widely addressed in other works. Our goal here is firstly focussed on the possibility of extraction of the FSP parameters.

We improve the robustness of the objective function by using a robust norm instead of the Euclidean norm in order to reject outliers. We use the Geman and McClure robust error norm [18] parameterized by a threshold σ :

$$\rho(X,\sigma) = \frac{\|X\|^2}{\sigma + \|X\|^2}. \tag{23}$$

The objective function to minimize is now:

$$E(p) = \frac{1}{n}\sum_{i=1}^{n} \rho(e_i(p),\sigma) \tag{24}$$

4.2 Levenberg-Marquardt optimization

The Levenberg-Marquaardt optimization can solve a non-linear least square minimization and therefore is well suited to model-based tracking by texture alignment as formulated above [7, 10]. The Levenberg-Marquardt algorithm requires the first and second derivative of the function to minimize with respect to every parameter, *i.e.,* the 3 rotations, the 3 translation and the 4 FSP parameters.

We introduce :

$$\psi(x,\sigma) = \frac{x}{x+\sigma} \tag{25}$$

allowing to express $\rho(X,\sigma)$ as :

$$\rho(X,\sigma) = \psi(\|X\|^2) \tag{26}$$

The first and second derivatives of the objective function are :

$$\frac{1}{2}\frac{\partial E}{\partial p_j} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \psi(\|e_i(p)\|^2,\sigma)}{\partial x}(\frac{\partial \hat{I}(p)}{\partial p_j}\cdot e_i(p)) \tag{27}$$

$$\begin{aligned}\frac{1}{2}\frac{\partial^2 E}{\partial p_j \partial p_k} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \psi(\|e_i(p)\|^2,\sigma)}{\partial x^2}(\frac{\partial \hat{I}(p)}{\partial p_j}\cdot e_i(p))(\frac{\partial \hat{I}(p)}{\partial p_k}\cdot e_i(p)) + \\ + \frac{\partial \psi(\|e_i(p)\|^2,\sigma)}{\partial x}(\frac{\partial^2 \hat{I}(p)}{\partial p_j \partial p_k}e_i(p) + (\frac{\partial \hat{I}(p)}{\partial p_j})\cdot(\frac{\partial \hat{I}(p)}{\partial p_k}))\end{aligned} \tag{28}$$

The partial $\frac{\partial \hat{I}(p)}{\partial p_j}$ with respect to a particular model parameter p_j is approximated by perturbing the parameter by a small δ, warping that model, and then measuring the resulting change in the residual error. The hardware texture mapping capability is very valuable for gradient calculations here [10]. As mentioned in [29], the second derivatives of the image can be omitted in the formulation of the Hessian.

The optimal value of the parameters p is then up-dated by iteratively taking the step δp that solves the linear equation formed from the approximation of the Hessian $\mathbf{H}$ and the gradient $\mathbf{g}$:

$$\mathbf{H} = \frac{1}{2}[\frac{\partial^2 E}{\partial p_j \partial p_k}]_{j=1\ldots10,k=1\ldots10}. \tag{29}$$

$$\mathbf{g} = [\frac{1}{2}\frac{\partial E}{\partial p_j}]_{j=1\ldots10}. \tag{30}$$

$$(\mathbf{H} + \lambda \mathbf{1})\delta p = -g, \tag{31}$$

The λ coefficient in Equation (31) serves to stabilize the inversion of the Hessian matrix when it is rank deficient.

5 Multi-texturing

Using a single texture for tracking makes an erroneous assumption as it links the lighting conditions and the geometric deformation of the model for one position only : as the model geometry is changing, the surface normals are changing as well but the lighting appearance is not. It results in unrealistic location of the lighting distribution, which could set the global minima reached by the optimization algorithm at a wrong configuration. We propose here an approach based on textures blending. In addition to cope with the lighting problem, textures blending allowed us to simulate small detail of skin deformation such as wrinkles, which are naturally appearing in the production of speech movements but are not geometrically represented by the sparse geometric mesh.

5.1 Linear alpha-blending formulation

For a given subject, six reference textures are chosen, corresponding to [a], [i], [u], [p] in [apa] and [f] in [afa]. The 3D model is aligned onto the images using the same procedure of features selection described in Section 3. Once the 3D mesh is aligned onto each image, the textures are blended into a linear class [7, 5].

$$\omega_k(a) = e^{-\lambda_k \sum_{i=1}^{n} \|X_i(a) - X_{i,k}\|^2} \tag{32}$$

$$\hat{\omega_k}(a) = \frac{\omega_k(a)}{\sum_{l=1}^{m} \omega_l(a)} \tag{33}$$

$$\hat{I}(p) = \sum_{k}^{m} \hat{\omega_k}(a) I_k \tag{34}$$

where $\sum_{i=1}^{n} \|X_i(a) - X_{i,k}\|^2$ is the sum of the Euclidean distances over all the n vertices between the current shape $X(a)$ and the reference shape X_k. The coefficient λ_k have been experimentally evaluated.

In the tracking procedure, the multi-texture showed an improvement in the detection of rounded shapes (“q[U]ick, br[OW]n”) by creating stronger minima of the error function at high value of FSP2 parameters, which corresponds to the rounding of the lips that occurs for this class of vowels.

6 Experiments and Results

6.1 FSP parameters extraction

We have tested the tracking of four different subjects uttering the same sentence: “That quick brown fox jumped over the lazy dog.” The subjects have been filmed with frontal faces with a stable lighting. The figure 4 shows the results for four subjects. The figure 6 displays the overall quality of the tracking for 6 important frames in the sequence of one subject.

The goal of the FSP description is first to provide a model that constrains the variation of the facial motion to a subspace specific to speech gesture. This approach allows us to introduce robustness into

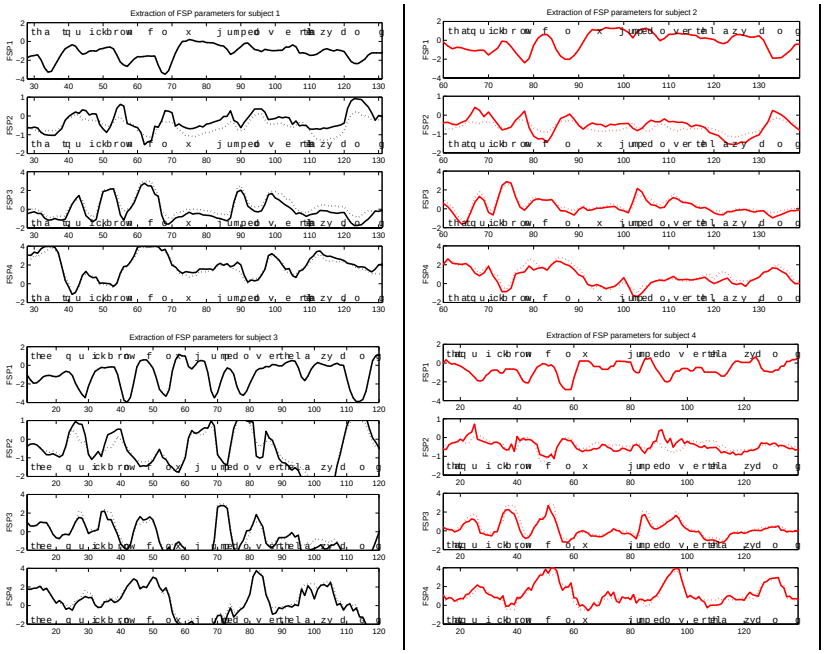

Figure 4: Results of tracking from video the four FSP for two subjects, uttering the same sentence. The dashed line correspond to the tracking with one single texture. The plain line correspond to the tracking with a multi-textures blending.

Figure 5: Example of textures blending. On the left, the original image. On the center, the result of the 3D model aligned on this shape, textured with the image corresponding to the rest position. On the right, the effect of texture blending, showing that the required wrinkles appear on the surface of the lips, while the texture corresponding to the face in a rest position has a low blending coefficient.

the parameter estimation. The risk of a drift and loss of the tracker is very low as the model remains on the constrained subspace.

In addition to this robustness in tracking, the claim of this approach is to bring a motion description that is independent of the speaker. All our subjects exhibited a stable behavior when we compared phonetically meaningful information across different speakers. It can be observed from the figure 4.

- the first FSP (opening of the jaw) shows a drop on the pronunciation of the [o] in “fox” and “dog”, corresponding to the jaw opening.
- the second FSP (lips rounding) reaches maxima on the expected rounded vowels [u] and [o]. The

Figure 6: The results at 6 key frames for one subject. Top to bottom: original sequence, textured model, wire-frame on top of subject and still picture animation from the data of this subject.

use of multiple textures (plain line) allows a better detection of these vowels, compared to the single texture case (dashed lines)

- the third FSP (lips closure) appears for the pronunciation of stop consonants ([b] in “brown” and [p] in “jumped”). This parameters shows high values for [f] and [v] consonants as well. However, these consonants are discriminated form the [p] and [b] consonants thanks to the fourth FSP (lips raising).

6.2 Application to facial animation

The texture-based model used for the tracking by alignment on images provides a photo-realistic facial animation solution (Figure 6). In addition to morphing different shapes, the blending of different views would allow a better full 3D representation as in [7]. Rendering of the teeth and tongue will have to be added for a complete photo-realistic perspective. As an extension of this, using the same procedure for aligning the model on a face shape, we have used the extracted FSP from video to animate the picture showed in Figure 6. As the FSP encodes natural gesture of speech production, this results in possibilities of high quality facial animation from one image only.

Finally, we have implemented an animation of 3D NURBS-based characters from the FSP parameters extracted from video. Instead of defining a lip and face shape for each phonemes like in a traditional lisynching process, we use a shape corresponding to the extreme variation of each FSP. Figure 7 shows an example of morphing targets suggested for the animation from the FSP extracted from video. It turned out to be more intuitive to design morphing targets with respects to extreme gestures (maxi-

mum opening of the jaw, maximum rounding of the lips, etc.), than the viseme approach in traditional lipsynch approach, where each speech sound must be assigned a corresponding shape.

7 Conclusions

The relationship of our coding to phonetic description is promising. We feel that such coding via FSP could be used as well to provide efficient visual cues for audio-visual speech recognition and bring robustness to the automatic speech recognition system. The main advantage of our FSP coding for video tracking relies on the fact that it constrains the complex behavior of facial movement of speech to only 4 degrees of freedom. However, this modelling currently does not cover motion variation due to expression that could occur with speaking. One of the natural extension of this work will be to investigate how the FSP coding could be extended to cope with facial expression and still preserves a stable behavior for tracking from video.

References

[1] V. Blanz, T. Vetter, "A Morphable Model for the Synthesis of 3D Faces", in *SIGGRAPH'99 Conference Proceedings*, pp. 187-194, 1999.

[2] M. Brand. "Voice Puppetry." *In Proceedings of ACM SIGGRAPH 1999*, pp 21-28, August 1999.

[3] C. Bregler, M. Covell, and M. Slaney. "Video Rewrite: Driving visual speech with audio". In Proc. ACM SIGGRAPH '97,1997.

[4] R. Campbell, B. Dodd, and D. Burnham, (Eds.) *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-visual Speech.* Psychology Press Ltd., East Sussex, UK. 1998.

[5] T. Cootes, G. Edwards, C. Taylor, "Active Appearance Models", *Proc. of the ECCV'98, Vol. 2, pp.484-498, 1998.*

[6] D. DeCarlo, D. Metaxas, "Optical Flow Constrains on Deformable Models with Applications to Face Tracking", in *International Journal of Computer Vision*, 2000.

[7] F. Pighin, R. Szeliski, D. H. Salesin, "Resynthesizing Facial Animation through 3D Model-Based Tracking", in *Proc. of the ICCV*, 1999.

[8] L. Reveret, C. Benoit, "A New 3D Lip Model for Analysis and Synthesis of Lip Motion in Speech Production", in *Proc. of the Second ESCA Workshop on Audio-Visual Speech Processing, AVSP'98, Terrigal, Australia, Dec. 4-6*, 1998.

[9] L. Reveret, G. Bailly, P. Badin, "MOTHER: A new generation of talking heads providing a flexible articulatory control for video-realistic speech animation", in *Proc. of the 6th Int. Conference on Spoken Language Processing, ICSLP'2000*, Beijing, China, Oct. 16-20, 2000.

[10] S. Sclaroff, J. Isidoro, "Active Blobs", in *Proc. of the ICCV*, pp. 1146-1153, Mumbai, India, January, 1998.

[11] S. Basu, N. Oliver, A. Pentland, "3D Modeling and Tracking of Human Lip Motions", *Proc. of ICCV'98*, Bombay, India, Jan. 4-7, 1998

[12] S. Basu, S. and I. Essa. "Motion Regularization for Model-based Head Tracking.", *Proceedings of International Conference on Pattern Recognition*, Vienna, Austria, August 1996

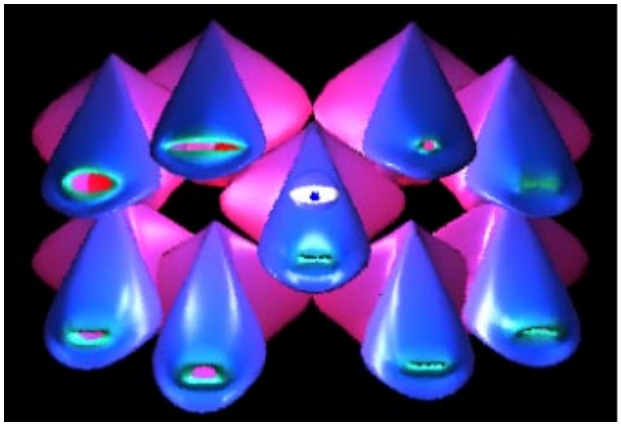

Figure 7: Morphing targets associated to the extremes positive and negative variations of the four FSP for a NURBS-based computer graphic character.

[13] M. J. Black and Y. Yacoob. Tracking and recognizing facial expressions in image sequences, using local parameterized models of image motion. Technical Report CAR-TR-756, Center of Automation Research, University of Maryland, College Park, 1995.
[14] S. Dupont and J. Luettin. "Audio-Visual Speech Modelling for Continuous Speech Recognition", In *IEEE Transactions on Multimedia* Vol2, No 3, p141-151, 2000.
[15] P. Eisert and B. Girod. Analyzing facial expression for virtual conferencing. *IEEE Computer Graphics & Applications*, 18(5), September – October 1998. ISSN 0272-1716.
[16] P. Ekman and W. V. Friesen. *Facial Action Coding System*. Consulting Psychologists Press Inc., 577 College Avenue, Palo Alto, California 94306, 1978.
[17] I. Essa and A. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 19(7):757–763, 1997.
[18] S. Geman, D.E. McClure, "Statistical methods for tomographic image reconstruction", *Bull. Int. Statistic Institute*, LII-4, 5-21, 1987.
[19] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice Hall, 1998.
[20] T.S. Jebarra, A. Pentland, "Parametrized Structure from Motion for 3D Adaptive Feedback Tracking of Faces", in *Proc. of CVPR'96*, CVPR, 1996.
[21] M. La Cacia, S. Sclaroff, V. Athitsos, "Fast, Reliable Head Tracking under Varying Illumination: An Approach based on Registration of Texture-Mapped 3D Models", in *IEEE Transactions on PAMI*, Vol. 22, No. 4, pp. 322-336, April, 2000.
[22] J. J. Lien, T. Kanade, J. F. Cohn, C. C. Li, and A. J. Zlochower. Subtly different facial expression recognition and expression intensity estimation. In *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition 1998*, pages 853–859, 1998.
[23] K. Mase. Recognition of facial expressions for optical flow. *IEICE Transactions, Special Issue on Computer Vision and its Applications*, E 74(10), 1991.
[24] K. Mase and A. Pentland. Lipreading by optical flow. *Systems and Computers*, 22(6):67–76, 1991.
[25] D. W. Massaro, M. M. Cohen, J. Beskow, and R. A. Cole, "Developing and evaluating conversational agents." In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill (Eds.) *Embodied conversational agents.* Cambridge, MA: MIT Press, 2000.
[26] F. I. Parke and K. Waters. *Computer Facial Animation*. AK Peters, 1996.
[27] C. Pelachaud, N. Badler, and M. Viaud. Final Report to NSF of the Standards for Facial Animation Workshop. Technical report, National Science Foundation, University of Pennsylvania, Philadelphia, PA 19104-6389, 1994. Available from http://www.cis.upenn.edu/.
[28] E. Petajan. Automatic lipreading to enhance speech recognition. In *Computer Vision and Pattern Recognition Conference*. IEEE Computer Society, 1985.
[29] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, "Numerical Recipes in C: The Art of Scientific Computing", Cambridge University Press, second edition, 1992.
[30] A. Schödl, A. Haro, I. Essa, "Head Tracking Using a Textured Polygonal Model", in *Proc. of the 1998 Workshop on Perceptual User Interfaces*, 1998.
[31] A. Smolic, B. Makai, and T. Sikora. Real-time estimation of long-term 3-d motion parameters for snhc face animation and model-based coding applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(2):255, March 1999.
[32] D. G. Stork and M. E. Hennecke (Eds.). *Speechreading by Humans and Machines.* Models, Systems, and Applications Series: NATO ASI Series, Vol. 0, Springer Books, 1996.
[33] A. M. Tekalp and J. Ostermann, "Face and 2-D mesh animation in MPEG-4," *Signal Processing: Image Communication, Special Issue on MPEG-4*, vol. 15, pp. 387-421, Jan. 2000.
[34] D. Terzopoulos, K. Waters, "Analysis and Synthesis of Facial Image Sequences Using Physical and Anatomical Models", in *IEEE Transactions on PAMI*, Vol. 15, no. 6, pp. 569-579, 1993.
[35] H. Tao, H.H. Chen, W. Wu, and T.S. Huang. "Compression of mpeg-4 facial animation parameters for transmission of talking heads". *IEEE Transactions on Circuits and Systems for Video Technology*, 9(2):264, March 1999.
[36] K. Waters and T. Levergood, "An automatic lip-synchronization algorithm for synthetic faces", *Proc. of the Multimedia Conference*, pages 149-156, San Francisco, California, September 1994. ACM